JLIS.it

Cataloguing in the open: the disintegration and distribution of the record

Martin Malmsten

Background

LIBRIS, the Swedish Union Catalogue saw first light in 1972 with the mission to rationalize work in libraries through the use of information technology. To this end a cooperative cataloguing environment was created where each member library would describe it's collection. The core idea, much in the spirit of the times, being that every library contribute what was unique to them, but still gain access the sum of descriptions from all member libraries. Access to the database was restricted to member libraries with access to a 3270 terminal and a modem connection the the LIBRIS mainframe. In 1997 a service was created to search and display the data on the web, in effect making the information in LIBRIS accessible to anyone, anywhere provided they had access to the internet. At around the same time a Z39.50[1] target was made available making it possible for copy-cataloging clients to remotely search and download data from LIBRIS. This was a important also since it allowed machines to access information about the collections described in

[1] http://www.loc.gov/z3950/agency.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5512

LIBRIS. Though Z39.50 was deemed sufficient for copy cataloguing, assuming the same format is used, it does lack in other areas such as userfriendlyness and has proven a problem when trying to connect to others outside the library community.

In 2006 a new catalogue frontend, a "next generation catalogue", was launched and with it a number of Application Programming Interfaces (APIs) meant to make it easier for anyone to create services built on top of LIBRIS' data. A consequence of this new catalogue was that a URI, or a URL rather, was minted for each bibliographic record making it possible for the system itself and external applications to link to individual records in a persistent and easy way. While this was all good and well given that you knew (a lot) about the APIs and could understand the formats provided (MARC21, DC, MODS, etc.), everything was still centered around descriptions contained within records. Also, few links to resources outside LIBRIS were present in the records once they had been found, even tough links to sites such as Google Books were provided to a human user of the frontend.

In 2008 LIBRIS as a whole was released as linked data (Malmsten), including authority data describing persons, organizations and subject headings. Links to external resources such as those described by Library of Congress Subject Headings (LCSH), Wikipedia and the Virtual International Authority File (VIAF) were added and in an instant LIBRIS was part of quickly expanding graph of metadata generated by a number of entities, mostly outside the GLAM sector. This move garnered a lot of interest especially from other government entities and other organizations that wanted to either link to or download parts of the authority data. This is unsurprising since an identifier for, e.g, a famous author is useful both for libraries, archives and other cultural heritage institutions.

Cataloguing in the open

Starting September 2011 the National Bibliography and Swedish Authority file, two subsets of the LIBRIS database, are made available in the same format that they are being created (MARC21). This decision to not only expose Resource Description Framework (RDF)/linked data derived from the records, but also the records in their original form is a strategic one. By doing so anyone can see, evaluate, reference and ultimately contribute to the work done by the National Library, the assumption being that visibility and openness will in the end lead to higher quality data.

To avoid any restrictions when it comes to re-use the National Library has chosen[2] CC0[3] for the National Bibliography and the Authority File, effectively putting the datasets in the public domain. The only exception is MARC field 667 (Nonpublic note) which is filtered out due to reasons of personal integrity. The license was chosen because we see a problem with attribution licenses such as ODB-BY and CC-BY when it comes to re-use of data over time, for example so called Cattribution stacking". The goal is to release to whole dataset in the original format with CC0, though since some records or part of records in LIBRIS originates from a number of other organizations (LC, BNB, DB, OCLC, etc), this will take some time. Anyone wishing to access the data can do so in two ways: either through Atom feeds and/or using the OAI-PMH protocol. The feeds are, essentially, getting the data from the cataloguing system in real-time. This means that anyone can get changes made to the Swedish Authority File within mere milliseconds of the change being made. The work of the cataloguers is being made available essentially as they type. As an aside the choice not to provide complete files for download (a

[2] http://librisbloggen.kb.se/2011/09/21/swedish-national-bibliography-and-authority-data-released-with-open-license.

[3] http://creativecommons.org/choose/zero.

"dump") of the data is to signal that the dataset is live, whereas a dump is essentially stale and/or obsolete in the same instant you download it. However, both Atom and OAI-PMH can easily be used to download the whole dataset, so the distinction is perhaps somewhat academic.

Consuming linked data

To actually reap the benefits of linked data, however, we must also use it as an integral part of our systems, not only expose it. This has a number of interesting consequences. First, since linked data allows us to relate to any data, wherever it is created, the distinction between internal and external datasets disappear. This has a profound impact on systems design since you have to rely on protocols normally used for external datasets internally. Secondly, the matter of control then becomes a matter of trust rather than technology. If you cannot control the information you must decide who to trust, and perhaps even cooperate with them though that is surely a small price to pay for a world of data. Often those we trust will also be consumers of the data we produce. Thirdly, as more and more information in our records relate to some resource outside of our control, be it a person in Wikipedia or a subject heading in id.loc.gov, the idea of a record becomes somewhat less interesting. A lot of what was the record is then actually controlled by descriptions that live elsewhere, in datasets merely linked to rather than owned.

Seamlessness

Often we do not need to aggregate data produced by others, but rather react to the fact that something that we link to has chan-

ged. Again this makes complete downloads of datasets at discrete intervals problematic since

1. the purpose a change may not be apparent and
2. having multiple batch imports that relate to same data will definitely cause problems.

A situation where synchronisation of datasets is done through download of the whole dataset simply does not scale. A goal for a truly linked system must be to be able to signal changes, or information about changes, seamlessly in near real-time to interested parties. While creating feeds makes it possible for clients to ask for updates, more often than not the answer will be that no, no updates have been made. This makes for a very inefficient system where a lot of requests have to be made to a ensure that the datasets are in sync. There are at least two efforts that deal with this issue: pubsubhubbub[4] and ResourceSync[5]. By using hubs to which a publisher can signal a change and a consumer can subscribe, an efficient network is created through which information can be sent.

Conclusion

Actually using linked data, as opposed to only exposing it, somewhat removes the distinction between internal and external datasets. Control becomes at matter of trust, not technology. The record disintegrates as the data becomes distributed.

[4] https://code.google.com/p/pubsubhubbub.
[5] http://www.niso.org/workrooms/resourcesync.

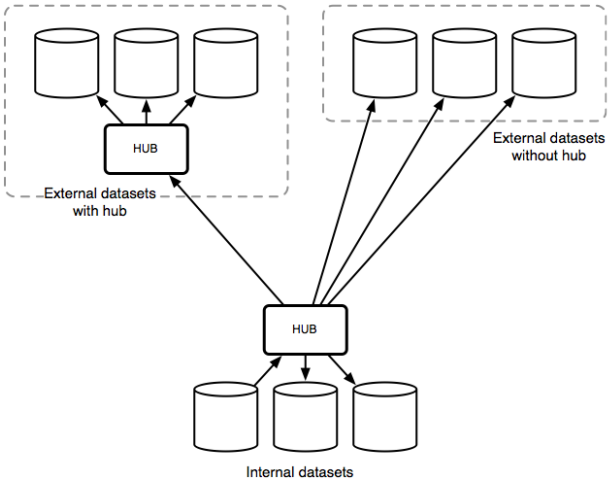

Figura 1: A change in one dataset is propagated through hubs, both to internal and external datasets

Riferimenti bibliografici

Malmsten, Martin. «Making a Library Catalogue Part of the Semantic Web». *International Conference on Dublin Core and Metadata Applications, DC-2008–Berlin Proceedings*. 2008. (Cit. a p. 418).

MARTIN MALMSTEN, National Library of Sweden.
martin.malmsten@kb.se

Malmsten, M. "Cataloguing in the open: the disintegration and distribution of the record". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5512. DOI: 10.4403/jlis.it-5512. Web.

ABSTRACT: As part of a strategic investment in openness the Swedish National Library has released the National Bibliography and accompanying authority file as open data with a Creative Commons Zero license effectively putting it in the public domain. The data has been available as linked open data since 2008 but is now also released in its original, complete form making it fit for re-use by other library systems. An important principle of linked data is to link out the other datasets. However, as data becomes more interconnected and distributed the need for ways to track and respond to changes in other datasets, even ones outside our area of control, becomes bigger. The issue of who to trust of course becomes vitally important. This paper details the motivation behind the release as well as the technology used to support it. Also, a consequence of exposing and using linked data is that the idea of the record as a self contained and delimited entity starts to fall apart.

KEYWORDS: Library linked data

Submission: 2012-04-25
Accettazione: 2012-08-31
Pubblicazione: 2013-01-15